

Data Organization

Welcome back to part two on this introduction to data science. Today what we're going to be talking about is data organization. So one of the biggest pieces of being a data scientist is organizing data. So I have this fake infographic here. 62% of time is spent organizing data. But really it's just the majority of time. So you're gonna get these data from different sources and you really wanna try to understand what they are. But also, data aren't coming in cleanly a lot of times. So, an example is a spreadsheet. So, maybe a whole bunch of people at your company have access to a certain spreadsheet, and the way they enter data are different. So, maybe some people use text in a column that should be all numbers, or maybe they leave it blank every once in a while. And so, all of these things can affect your ability to analyze the data effectively. And so, you have to go through and kind of munge that data and clean it up to make sure that you can actually use it. And as the data sets get larger and larger, the amount of time that you spend doing that and kind of fixing things that are not right becomes greater and greater. So that's one really big aspect of kind of being a data scientist is just sitting and working with data sets so you can get it into a format that you can actually use. Setting up your columns and rows and making sure that you have potentially multiple rows that are kind of describing things, so you can organize it better later on. All those things are part of this cleaning up and organizing data. In addition, not all data types come in this sort of tabular form, right? So data scientists are dealing with live stream video, as I've mentioned, and how do you deal with this information that's coming in? How do you deal with this audio signal that's coming in? So those data are totally different than what you think of or we generally think of as an Excel table. But there are important types of data and you need to figure out how to deal with them. So again, just to kind of reiterate on this point. Organizing data is a huge part of being a data scientist. And it's kinda fun once you kinda get used to it, but it takes a little bit of time to kinda get your hands around data and really feel comfortable and being fun. But let's just talk about kind of how I break apart organizing data. And I'm gonna do this for each of these three pieces of

organizing data, visualizing data, and analyzing data.
I kind of split it into three different levels.
Some examples, and so a great track to becoming a data scientist is kind of work through each level by itself.
But that's not the only track.
So let's just go ahead and start and talking about the kind of three levels of organizing data.
So the first one is kind of this level one data.
And these are going to be like Excel tables, so these are things you'd see in a small business. Where your company has information about it.
In this one table, and it's kinda pretty organized structure.
So maybe it's a MySQL table that you're pulling in from MySQL or some other database as well.
But it's in really quite structured and quite tabular.
The second level is gonna be the sort of more complicated data types. So Genomics is the example that we're gonna talk about today.
But it's not just genomics.
It's basically when you're pulling in data from different locations and trying to integrate it.
So maybe you are working at mid-size company that's dealing with shipping.
And you wanted to look at shipping over time and you noticed kind of these weird trends that there were these downtime in shipping.
And so maybe, as a data scientist, you're wanting to pull in weather information at the same time.
And so your Excel table that you're using for your company doesn't have weather information so you're having to go and get that data from somewhere else and bring it in.
And so you're doing these mashups of data.
And so kind of Level II is being really comfortable mashing up that data and bring it in to these more complicated formats.
And then Level III is kind of this next step.
And it's not necessarily a step in the Level III stuff, it's dealing with data types that are just non-uniform.
So these data types, they're just things that you aren't used to dealing with and figuring how to deal with those.
And so those are kinda my three levels of being a data scientist and organizing data.
So let's just talk about level one and two.
Level one first, and then we'll move into level two, and then into level three.
And throughout this course, once we get into the demos, we're gonna be focusing on kind of this level one, maybe a little bit into the level two stuff.
I think some of the other courses that are out there, so

Microsoft has a great certificate program that really talks about moving from level one to level three on some of these things. And so, if you're wanting to explore more after this set of courses, or you're just like, I love data science, I just wanna jump right into that, that's a great place to start, so cool, all right. So, level one data. I think most, everybody has seen this sort of data. It's a sort of common tabular format where you have information in columns and rows. Each row, generally, is one piece of information about a product or a date or some sort of things. So, here's an example of maybe just a profit chart that you might see in a small company. So, as we kind of go through this, I wanna point out a few different things. So, you'll notice on line five here, this data, as we mentioned, isn't really clean, right? So we have all these numbers that are formatted in dollars and then we have this random string of 14. And so, these features of these Level I data types are generally, they're pretty humanly readable. So there's not like a million or 2 million or 100 million lines of different data. It's something that you could read through and scroll through yourselves. So maybe there's 100,000 of them, you could still do that yourself and do it in a reasonable amount of time. So humanly readable, usually fairly clean, so you're not having to change a lot of things, maybe a few things you're gonna have to go through and edit. And it's usually coming from a single source. And so, as I said, where you'd see this sort of thing is maybe small businesses, maybe in the biology scene, you'd see them FDA trials of well-designed experiments, that sort of thing. But having a simple data set doesn't really mean the analysis is automatically simple. So in fact, small data sets can also be harder to analyze in some ways because you don't understand what's going on in the data. There just aren't enough data points to really get your head around them. So that's kind of the general sense with level one data. What sort of tools would you wanna have as a Level I data organization, data scientist? Well obviously you wanna be able to use spreadsheets or deal with tabular data. There's some great tools for kind of organizing data, so Excel is great, Power BI is great.

You probably want to understand some basic SQL Query and different types of data formats that are tabular in nature but maybe not in format.

So comma separated value files for example is a really common data type format.

So these are formats where basically, instead of rows and columns, you have enter, generating a new row, and commas separating each column.

Cool, all right, so moving on to Level II data.

So level II data are data that may not be humanly readable, but are still parsable.

Here's an example of a gene sequence.

So, basically, just a whole bunch of As, Ts, Cs, and Gs.

An average one of us looks at this, they're gonna be like, okay, that's great.

[LAUGH] But

we can't really do anything with that in its current format.

We can do a few things, but just it's not very usable.

And so, a level two data scientist might be taking this data and applying some different structures to it.

So we know that genes can be turned into proteins and there's some rules about doing that, right?

And so basically DNA has code ons in it.

And those code ons can get turned into proteins through transcription and translation.

And so we know those rules.

We know how to do it.

And so this data set might turn into some sort of column and row, sort of level one data in the following manner.

So basically, all I did is counted the number of nucleotides in the sequence, and then converted into a protein, found the longest reading frame, and said, hey, that protein length is gonna be 566 amino acids.

So that's just an example of kind of taking multiple pieces of information, bringing them in, and then turning them into something that's a little bit more digestible.

And so, that's kind of the goal of the level II,

Aspect of data organization.

And so, sometimes it's humanly readable and you're bringing in multiple tables.

Other times, you might need some

domain expertise to kind of really get your hands into that data.

And where you're gonna see this sort of data are genomics, larger corporate entities, potentially the streaming data that's coming back with really clear formatting, but it's coming fairly quickly.

So all those things are kind of this level two data structure, where you're really getting and

organizing this stuff in a really tangible and parsable manner.
Cool, so again, some tools that you really wanna have kind of down when you're jumping into that level II and moving on to that level II. Again, you're gonna be using PowerBI probably, or Tableau, or some of these similar tools.
But you're also gonna be starting to dive into some sort of programming language.
So it's really hard to look at a million rows of data and bring over tables together.
And then clean that data up without some sort of statistical programming package.
So some of the ones that people are commonly using, Python with Pandas or Spark, R, you'll see some Java and you'll also see some functional programming and stuff with Scala.
So again, as you move from this level one to level two, you're gonna kinda move from this really clear, readable data to kind of this slightly more abstract level.
And so, being able to, kind of pull things in, and understand how you're pulling them in without really seeing all the data right there.
So, cool, so the skills that you wanna have, you definitely want to be able to mash up data.
You want some sort of data intuition, so when you look at some of your visualizations, you kind of know how you're going to potentially attack it later.
Because you can't do everything visually, because there's just too much information there potentially, or it's not humanly readable, you really need to be confident in your ability to programatically clean up that data.
So this is kind of data munging, and so you really wanna trusting over your skill set and be like, everywhere there is 0 I'm gonna convert that to a 42 because it makes sense to do that for this prior dataset, maybe we're to figure out what the meaning of the word is or something.
So those are kind of the skills that you wanna have.
And again, although I kind of wanna emphasize this point of data scientists being kind of data agnostic, you can bring in information from wherever and apply these really standardized best models and best attack methods to attack that data.
In some instances, data area, our subject area knowledge is actually advantageous.
So it's not saying to be ignored, but I do like this idea of even if it's just a theoretical idea of data scientists being data type or data source agnostic.
So, cool, so let's move on to Level III data organization.
So I'm not gonna talk too much about this, but Level III data organization is really when you're starting to get and

deal with non-uniform data.

So we talked about this idea of data exhaust in our first chapter.

So this idea that as we move through the world, we're doing all these things that are being picked up by our behavior.

So our searches,

maybe traffic cameras, maybe if you're using a smart fridge.

All these things are generating data and

there's so much data that's coming through that one of the strategies that people have used this like, who's gonna collect this stuff.

We don't care how we collect it,

how organized we are about collecting it.

We're just gonna grab it.

So by grabbing it and just having it there for us.

We need some way to deal with that and

figuring out how to restructure that data from this kind of unstructured format and bring it in to something that we can actually use is a big part of this Level III data organization.

In addition you're gonna be dealing with things that are just huge amounts of data.

Where you're basically getting so much information in that maybe your machine can't deal with all that data at one point, so you're gonna have to use distributed systems.

Or that data is coming in so quickly and you wanna get it back out in some sort of digested format, very quickly as well.

So real time integration, where you're trying to really optimize how you're organizing that data and

how you're spinning it back into different locations.

And so that's kind of the level three dataset, skillset that

I think you really wanna have down if you wanna be a level three data organization piece of the data scientists collection.

So what sort of tools would you want for being level three?

So again now we're moving into specifically into programming.

So you're going to be using Python, you're going to be using R,

Scala and you're really gonna be wanting to know how to deal with distributing systems.

So Hadoop, or Pig, or Hive are some of these just databases that can really, are set up to deal with big data sets.

And once you kind of have those down, and if you're able to take information, kind of from anywhere, and bring it in, you should feel confident that you're kind this Level III data organization piece.

And one of the things I wanna kinda just bring up one more time is a data scientist does not have to be Level III on every single thing.

So, some of its Level III on data organization is extremely valuable because you might have a level 3 data analysis out there whose

gonna take the data once you've generated it and just make some amazing conclusion off of it.

Science in general as progressed from
this idea of you're an individual can do everything
to this kind of collaborative environment, and
data organization or data scientists are no exception to that.
There's a bunch of really specialized skill sets.
And you could be great at all of them or
you could work as a team and be just phenomenal at one of them and
let your other team members kinda take over those other roles for you.
So don't be afraid to like, I love doing this one aspect of data
science and think that that's not valuable, cuz it is, right?
So, next we're gonna be talking about data visualization.